Data Analytics Found/Practicum

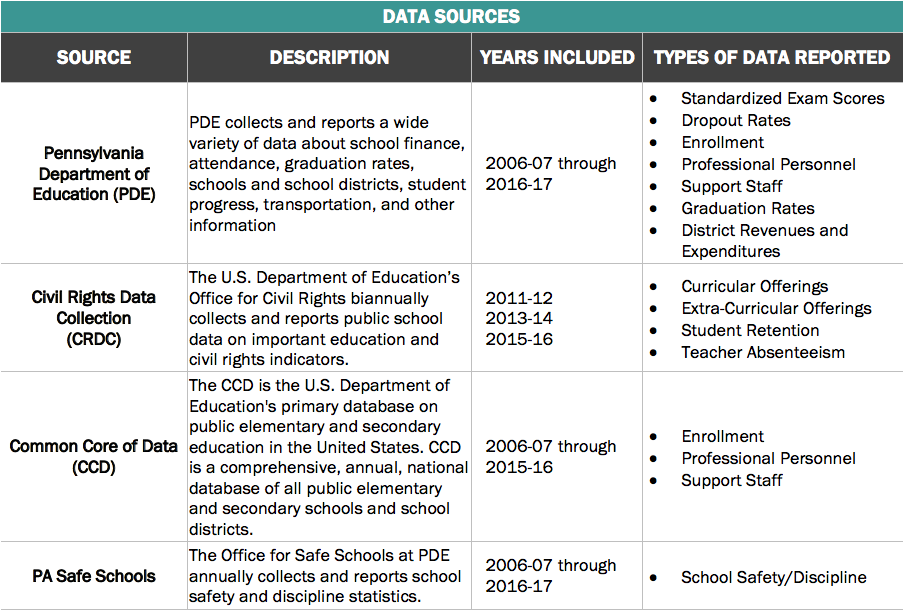
Austin Gerner

Dr. Benjamin Harvey

Exploratory Analysis 2015 - 2016 Pennsylvania Public High School Data

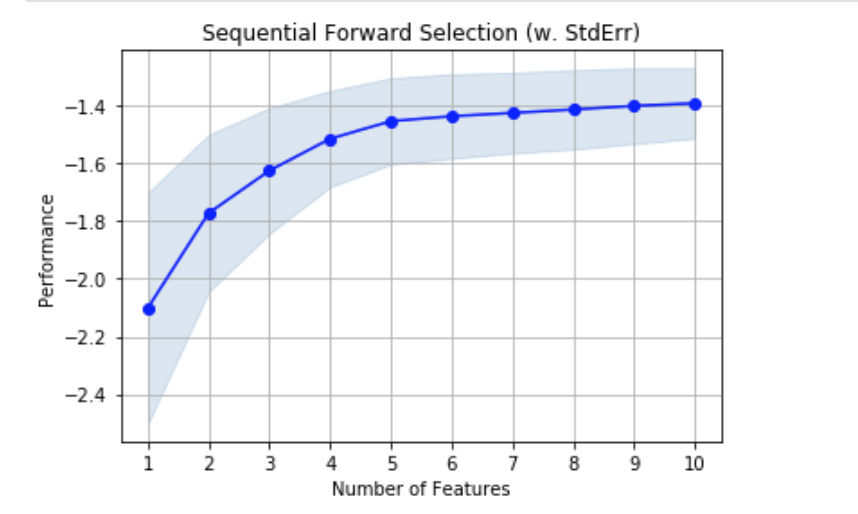
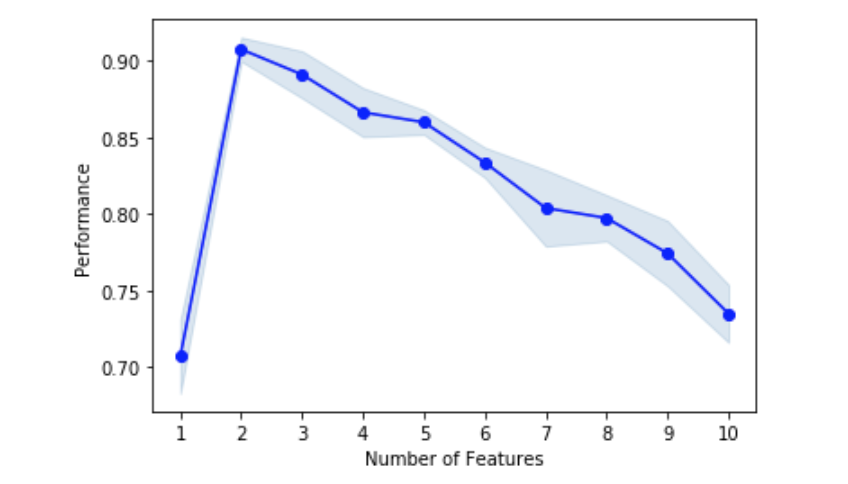
In today’s current job market it has become a standard for most companies to hire employees with a post secondary education background. The National Center for Education Statistics reported in 2018 that employee rates of 25 to 34 year olds with a highschool diploma had an employment rate of 72% [1]. The average salary of that 72% was around $37,000. Where as an individual with post secondary education saw their employment rates increase to 79%, and their average salary rise 8% to around $40,000[2]. With the devaluation of the high school diploma and the increase in competitiveness for college admissions these pressures force students and parents to start the college process earlier. Families are forced to properly prepare for post highschool[2]. Highschools have a huge responsibility to give higher quality education to give their students an edge in whatever post secondary route they decide to take. Parents need to be able to identify school districts that have strong programs, with proven results to give their children the best chance at being able to attain secondary education.

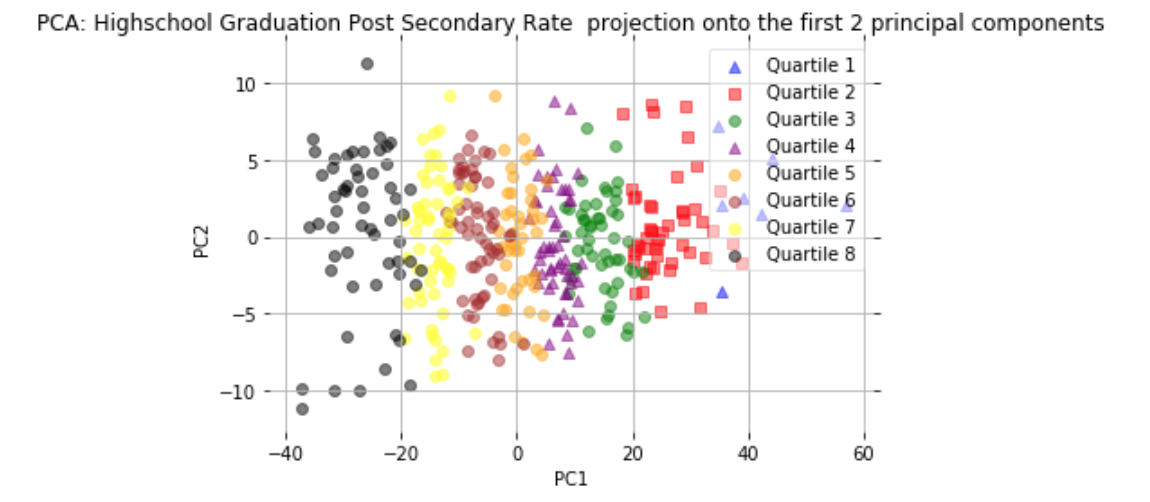
With the passing of the No Child Left Act in 2001, this legislation required schools to annually test their students and report on their findings[3]. Results from these tests showed that most students were struggling based on the education they were receiving in their highschools. Former chancellor of the District of Columbia Public Schools, Michelle Rhee spoke on this topic, “When we took control of this school district in 2007, 8 percent of the 8th graders were operating on grade level in mathematics—8 percent. And if you would have looked at the performance evaluations of the adults in the system at the same time, you would have seen that 95 percent of them were being rated as doing a good job.”[3] No longer could struggling students be ignored, when the data clearly was telling the country that there was an education crisis occurring. In 2009 the Obama administration spoke to these poor test scores with the creation of the Race to the Top program [4]. Through means of federal funding this program helped motivate districts to develop measurable standards, more effective testing, and creation or adoption to data systems allowing for better communication between parents, students, and teachers [4]. Reform programs like Race to the Top, as well as modifications to individual state laws created a massive influx of data across the country. This data allowed schools to make effective data-driven decisions and created an environment of innovation that allowed schools to be bolder and test new theories. With this consent stream of data it also allowed schools to reduce redundancy and cut programs that were failing [5]. These data sources are available to the public, which gives individuals, outside of academia, the opportunity to use current data science methodologies and tools to conduct their own analysis.

This projects scope focuses on public school districts in Pennsylvania, specifically highschools during the 2015-2016 school year. The first goal of the project was to explore all relevant data sources. This goal can be achieved by using current data identification and acquisition techniques, which merge into a cohesive dataset. Once accomplished, the next focus of this project was to conduct an analysis to find features that best predicted post secondary attainment rates. The main data sources leveraged for this project were from the Pennsylvania Department of Education (PDE), Civil Rights Data Collection (CRDC), Common Core of DATA (CCD), and PA Safe Schools [7]. These organizations collect reports on annual and biannual occurrences. As the chart below shows the data sets cover a wide range of features. After initial exploratory analysis on the datasets I classified each set into four categories: People, Outcomes, Opportunities, and Financial.The People’s dataset included features such as school enrollment breakdowns by grade, race, and gender. This was in addition to the number of students per school with specialized education plans, percent of families with lower income, staff breakdowns, and pupil/teacher ratio [7]. The Outcomes data set included features associated with test result breakdowns for standardized tests. This included the PSSAs, Keystones, SAT, ACT, AP Exams, internal school final grades, graduation rates, postsecondary rates, and attendance rates. Other notable features in the People data set was the occurrences of the various types of school suspensions. This included the racial and gender features of the individuals [7].

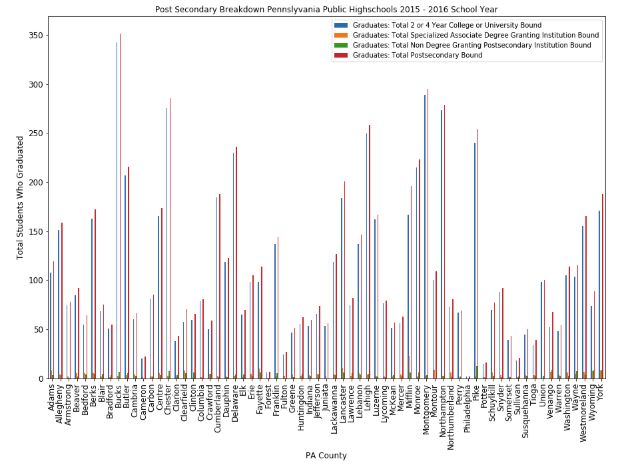
The Opportunities data set had records pertaining to the number of certain classes offered to students (i.e. algebra, geometry, calculus) [7]. Student involvement in activities like sports were broken into individual or team sports and then classified by gender. The number of students taking AP classes, and involved in gifted/talented programs were broken down by gender and race. The Financial data set had to do with all expenditures and revenues associated with each school district [7].

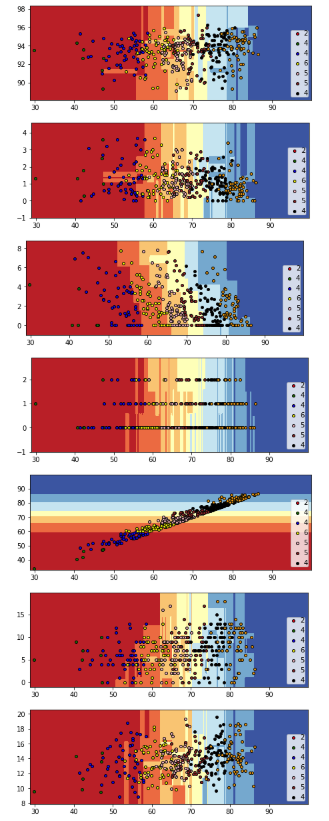
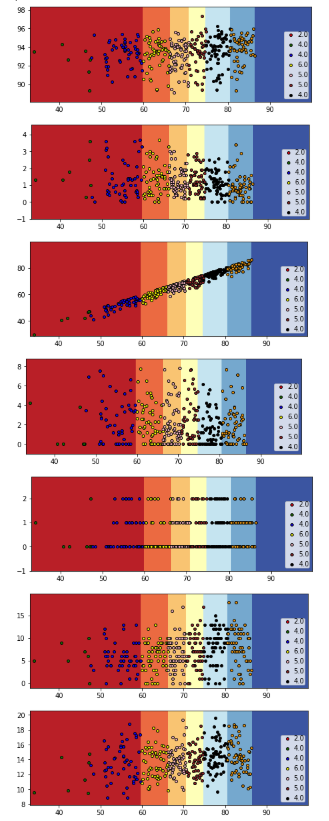
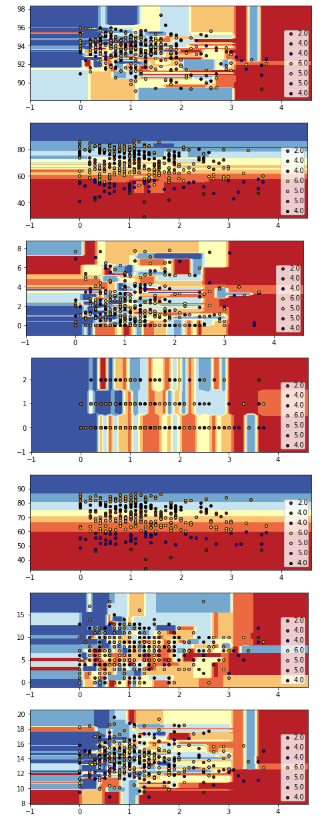
Once the features for each data set were identified and classified I started the process of data filtering, feature extraction, data validation & cleaning. This helped to narrow the data set from its original size of over 2700 total schools [7]. For this project the focus was highschools during a single year. The initial records removed were elementary and middle schools. Most of the data covered was from the 2006-2007 school year through 2016-2017. The reason the 2015-2016 year was used instead of the 2016-2017 is due to the fact that the 2016-2017 school year was not as robust. After the initial filtering was completed we found 601 highschools that fell into our scope [7].

Now that the data had gone through the initial round of cleaning and was put into one cohesive set, it allowed me to start researching into feature extraction. After, initially testing each feature individually through trial and error to post secondary rates, I started researching into more efficient methodologies to find meaningful features. First, I bucketed the post graduate rates into eight different buckets. Then after conducting research, decided to leverage sequential feature algorithms (SFA) to help reduce the dimensionality of the feature space [6]. Using sequential forward selection algorithms (SFSA) I was able to narrow down which features would have the greatest impact on predicting post secondary rates as well as the optimal number of features to avoid any unnecessary noise in the model. The figures above show the results from the SFSA analysis.

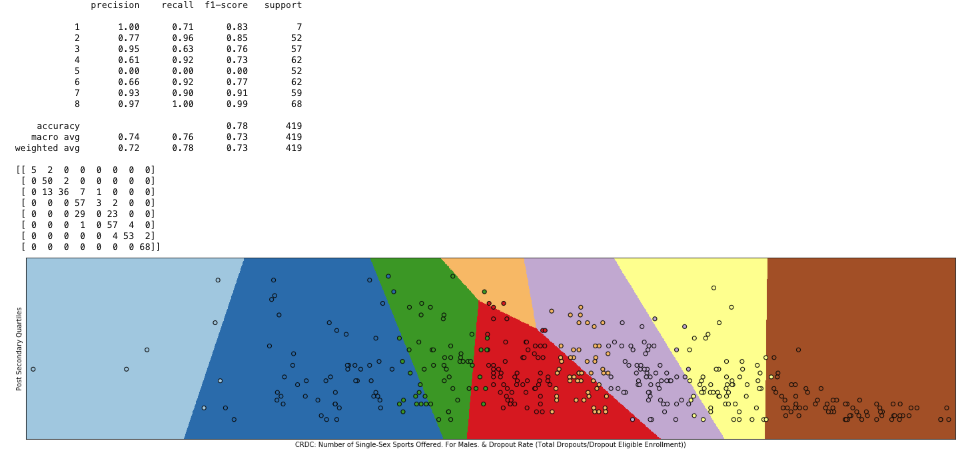
In this test we can see that the performance decreases as more features are added. Due to the fact that features like “Graduates: Percent College Bound” will have a higher performance feature than others leads to make up a larger portion of the post secondary attainment rate. I chose to use the top ten features selected, and then went into the process of filtering and cleaning the small feature set to remove any outliers in the data. This resulted in removing two additional features, and narrowing the amount of schools in the analysis to 419 public highschools. To confirm the data set was ready for a deeper analysis, I performed a PCA on the first 2 principal components. We can see from the chart below that the dimensionality of the data has been significantly reduced while maintaining some variation. 

The first part of my analysis had to do with having a better understanding of what made up the postsecondary totals along with the types of counties that received higher post secondary rates compared to others that year.

By looking at the chart below we can see that Graduates: Total 2 or 4 Year College or University Bound makes up most of the Post Secondary Totals. This means that most of the students who graduated in 2016 that did go on to a post secondary education eventually attended a 2 or 4 year college. The next part of my analysis had to do with researching the differences between counties that had higher post secondary rates compared to the entire data set. I found that counties that were above the average 66% post secondary attainment rate had a higher enrollment total by 56 kids. Counties with higher post secondary rates also had a 7% decrease in the number of families that were classified as low income. The pupil/teacher ratio in these schools saw an average drop of 1 teacher going from 15 students per teacher down to 14. The pupil/teacher ratio in these schools saw an average drop of 1 teacher to every 15 students. This eventually decreased to 1 teacher to every 14 students.

Once there was an understanding of the data I was ready to solve the main question of this analysis: to find the best features to predict higher post secondary attainment rates. The first step was to use the Decision Tree Classifier from the scikit-learn package to visualize the different feature combinations. Through visualization I was able to understand which features had better potential in predicting post secondary rates. From the results of the visualizations we can see the following three features stand out for predicting post secondary. Visualizations are represented by the following features from left to right: Number of Single-Sex Sports Offered. For Males., Graduates: Percent College Bound, and Dropout Rate (Total Dropouts/Dropout Eligible Enrollment). 

The next part of the analysis involved using a model that would best predict these features. This led me to use logistic regression as this type of regression is very useful for quantifying the impact each feature has on the outcome we are trying to predict (highest post secondary attainment rates). After training the model I found that the highest accurarcy for predicting Post Secondary Graduation Quartiles were from Number of Single-Sex Sports Offered. Males., and Dropout Rate (Total Dropouts/Dropout Eligible Enrollment). These two inputs predicts the Post Secondary Quartile accurately 78% of the time.

In conclusion, the captivating results make up a large portion of the post secondary totals due to the fact that Graduates: Percent College Bound makes up a large portion of the post secondary totals in itself. As a future study I would conduct a time series analysis on all of the available years data to see if the same features would be as influential as they were in this study. The financial data integrated into a model would also be beneficial. Once one finds the influential components (i.e. certain classes, afterschool programs, dropout rates) this will lead to higher post secondary graduation rates. Schools will then be able to review the budget and increase funding to eliminate programs that are not as effective versus those that are.

## References

1. Torpey, E. (2019). *Education pays : Career Outlook: U.S. Bureau of Labor Statistics*. [online] Bls.gov. Available at: https://www.bls.gov/careeroutlook/2019/data-on-display/education\_pays.htm [Accessed 10 Dec. 2019].
2. Nces.ed.gov. (2019). *Fast Facts: Employment rates of college graduates (561)*. [online] Available at: https://nces.ed.gov/fastfacts/display.asp?id=561 [Accessed 13 Dec. 2019].
3. Simon, R. (2019). *Data Was Supposed to Fix the U.S. Education System. Here’s Why It Hasn’t.*. [online] Harvard Business Review. Available at: https://hbr.org/2019/01/data-was-supposed-to-fix-the-u-s-education-system-heres-why-it-hasnt [Accessed 10 Dec. 2019].
4. US Department of Education (2009). *Race to the Top Program Executive Summary*. Washington, D.C, pp.1:5.
5. BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS. (2016). *International Journal of Modern Trends in Engineering & Research*, 3(11), pp.84-88.
6. Raschka, S. (2019). *Sequential Feature Selector - mlxtend*. [online] Rasbt.github.io. Available at: https://rasbt.github.io/mlxtend/user\_guide/feature\_selection/SequentialFeatureSelector/ [Accessed 7 Dec. 2019].
7. Research for Action. (2019). *Pennsylvania School Data Project - Research for Action*. [online] Available at: https://www.researchforaction.org/pa-school-data-project/ [Accessed 13 Dec. 2019].